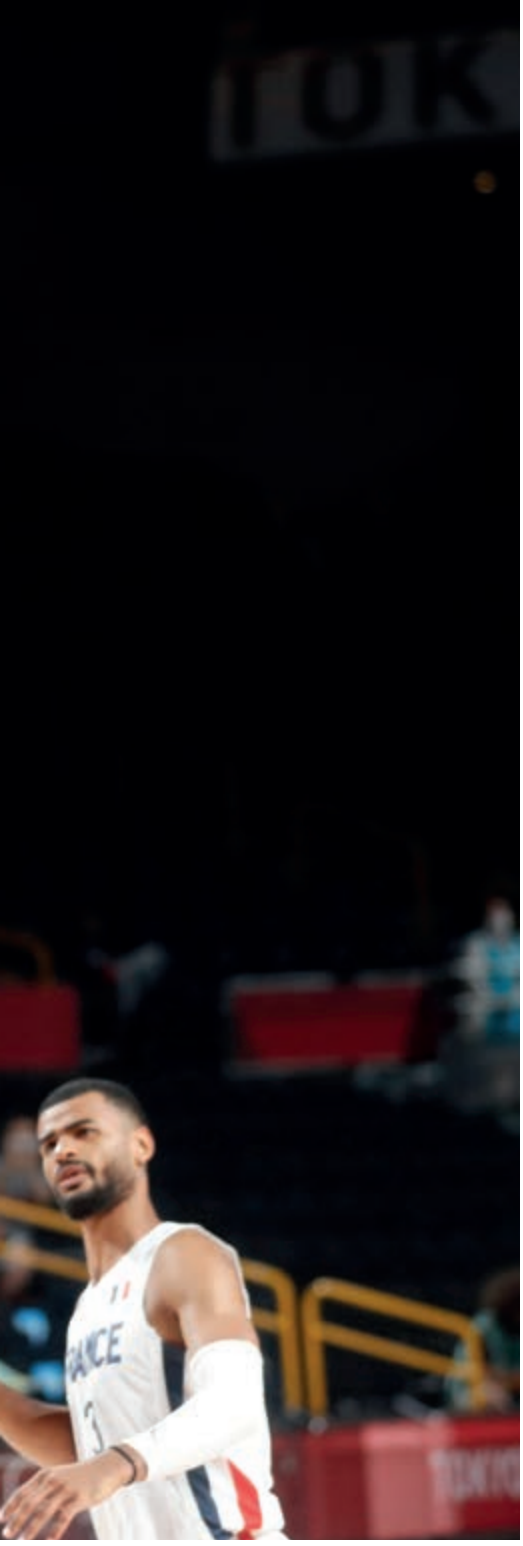
Slam dunk statistics

Despite other countries catching up, the USA remains a global basketball powerhouse and is favourite to win gold in basketball at the Paris 2024 Olympic Games. **Oliver Jack** and **Christophe Ley** use cluster analysis to help answer some of its fans' hottest topics of discussion



A Ricardo/Shutterstock.com

Thanks to the constant evolution of technology, most basketball leagues nowadays are able to provide a solid collection of statistics related to players and teams

Picture yourself in 1991 sitting courtside at the Chicago Bulls game. You hear the beautiful sound of the swish in the basketball net and the crowd erupts – they are witnessing the biggest show in town. The arena is buzzing with anticipation as Michael Jordan, the basketball phenomenon in red and black, takes to the court in the dying seconds of the game. The clock ticks down, and with a flick of the wrist, he sinks a game-winning fadeaway jumper, his signature move, writing another chapter in the history of the American National Basketball Association (NBA). It is more than just a shot, it is a glimpse into what will be cemented in history as the "classic era" of NBA basketball.

Fast forward to the present, where the three-point arc surrounding the basket has become a launch pad for a new breed of sharpshooters redefining basketball norms. Golden State Warriors hero Steph Curry pulls up from beyond the arc and sinks one of his numerous three-pointers – a symphony of precision and strategy that resonates through the "modern era". The contrast between these two moments is more than a shift in playing styles, it is a journey through time, showcasing the essence of the NBA's dynamic evolution.

Being a ball sport with a rich history, basketball has continuously grown in popularity over the last century. Currently, it is watched and enjoyed by millions of people worldwide on a daily basis. At the centre of it all lies the NBA, founded in 1946. It has witnessed an awe-inspiring journey from its humble beginnings to becoming a worldwide-televised sporting phenomenon. Meanwhile, the game itself and its strategies have evolved considerably over time, sometimes making it difficult to compare the new era of basketball to how it was decades ago. This has led to heated debates on critical topics such as who is the best player of all time, which players are underrated or overrated, and what were the golden years of this fascinating sport.

A sound use of modern statistical methods can help settle these discussions. Thanks to the constant evolution of technology, most basketball leagues nowadays are able to provide a solid collection of statistics related to players and teams. In particular, the NBA is at the forefront when it comes to gathering data. While certain information is kept private for the team's personal use, the vast majority of the collected data is accessible to the public. This has allowed more and more analysts and fans to form their own opinions based on stats, especially when making their assessments of players' and teams' performances. Despite the wealth of available data, the problem proves difficult because of the challenge we previously mentioned – today's game is vastly different from how it was decades ago, even more so for basketball than for other sports. The introduction of the three-point line in the 1979–1980 season, the frequent rule changes, and the continuously evolving playstyle are some of the more notable changes seen in recent years. As players evolve in athleticism and skill, coaches adapt their strategies to leverage these capabilities. As the league globalises, the influx of international talent incorporates diverse playing styles, further enriching the NBA's stylistic tapestry. Moreover, the advent of sports analytics has had a profound impact on the way the game is played.

Therefore, our goal was to obtain a data-driven understanding of the different eras in NBA basketball history. This would not only provide some concrete evidence to help settle the ongoing debates, but also provide invaluable insight about the sport and a better understanding of differences in player statistics between distinct eras. With the help of *k*-means clustering, our aim was to group NBA seasons into clusters with common traits and, consequently, identify the specific eras.

Cluster analysis is an unsupervised classification method, where the goal is to divide individual data points into groups (i.e. clusters) based on the similarity of their attributes. In the context of machine learning, "unsupervised learning" refers to a type of algorithm that learns patterns from non-labelled data, without any guidance on what specific structures should be found. In other words, data points that are grouped in the same cluster are supposed to share common characteristics with one another, while being markedly different to the data points in other clusters. A commonly used partitioning technique is the so-called *k-means clustering* method. In general, it can be divided into

Oliver Jack is in the first year of a master's in data science at the University of Luxembourg, and this paper is based on his bachelor thesis work.

Christophe Ley is associate professor of applied statistics at the University of Luxembourg.

Box 1: What is the coefficient of determination?

The coefficient of determination η^2 is defined as the ratio of the "between deviance" (*BD*) to the "total deviance" (*TD*) of data points. Let n be the number of data points grouped into k clusters, and let μ denote the average value of all data points. Then *BD* is defined as $BD = \sum_{i=1}^{k} (\mu_i - \mu)^2 n_i$, where μ_i is the average value of all data points in cluster i and n_i the number of data points in cluster i. On the other hand, *TD* is defined as $TD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu)^2$ where x_{ij} stands for the value of data point j in cluster i. It measures the overall variability of the data. The ratio *BD*/*TD* lies between 0 and 1, with a higher value indicating a better separation between clusters.

A commonly used approach for finding the optimal number of clusters is the so-called elbow method, which involves visually identifying a plotted point where the rate of increase in variance significantly clearly diminishes, resembling an "elbow"

two main steps: identifying the number of clusters to be defined and determining the respective clusters.

An important notion for the first step is coefficient of determination η^2 (also known as explained variance), which measures the clusterisation quality with respect to the number of clusters (see explanation in Box 1). After determining the number of clusters, k, which yields the highest quality, the method seeks the cluster subdivision that gives the best η^2 for that value k in a second step. This search consists of a repetitive algorithm that tends to optimise the partitioning of the data points into k distinct clusters. First of all, k cluster centres are randomly chosen, then each data point is assigned to the cluster whose centre is the closest according to a well-defined distance (typically the Euclidean distance). Next, the centre of each cluster is recalculated as the mean of all data points assigned to that cluster. The reassignment of the data points and the recomputation of the centres is then repeated until the centres become relatively stable.

In our case, the attributes considered will be a collection of basic basketball stats which we provide in Box 2, while the observations will simply be the seasons from 1980 to 2023 (prior to 1980, not all of these stats were available, due to the three-point-line not being introduced yet). With the help of the `kclustering` method of the BasketballAnalyzeR package in the software R (package accompanying the commendable book by Zuccolotto and Manisera[1]), the graph summarising the average coefficient of determination per number of clusters is obtained and shown in Figure 1.

The explained variance naturally improves with the number of clusters. A commonly used approach for finding the optimal number of clusters is the so-called elbow method, which involves visually identifying a plotted point where the rate of increase in variance clearly diminishes, resembling an "elbow". This technique aims to find the right balance between the number of clusters and the clarity of interpretation, determining whether an additional cluster is justified or not. Looking at Figure 1, one can recognise that the optimal number of clusters is 3, with an explained variance of 80.86%.

The next step of the clustering algorithm was executed by running the `kclustering` function once again, this time specifying the number of clusters. As a result, the following three clusters are obtained, each of which supposedly represents a different NBA era: 1980–1994 (first era), 1995–2015 (second era) and 2016–2023 (third era). A first observation is that there are no gaps between the seasons or eras, guaranteeing a smooth transition from one era to the next. It is noteworthy that this property occurred naturally, rather than being imposed by the initial modelling. Interestingly enough, this property of "smoothness" remains true if the number of clusters is increased to four (1980–1993, 1994–2005, 2006–2016, 2017–2023) or five (1980–1989, 1990–1994, 1995–2005, 2006–2016, 2017–2023). Nevertheless, we will stick to the three originally obtained eras for the remainder of the analysis. Although the second era is visibly the longest

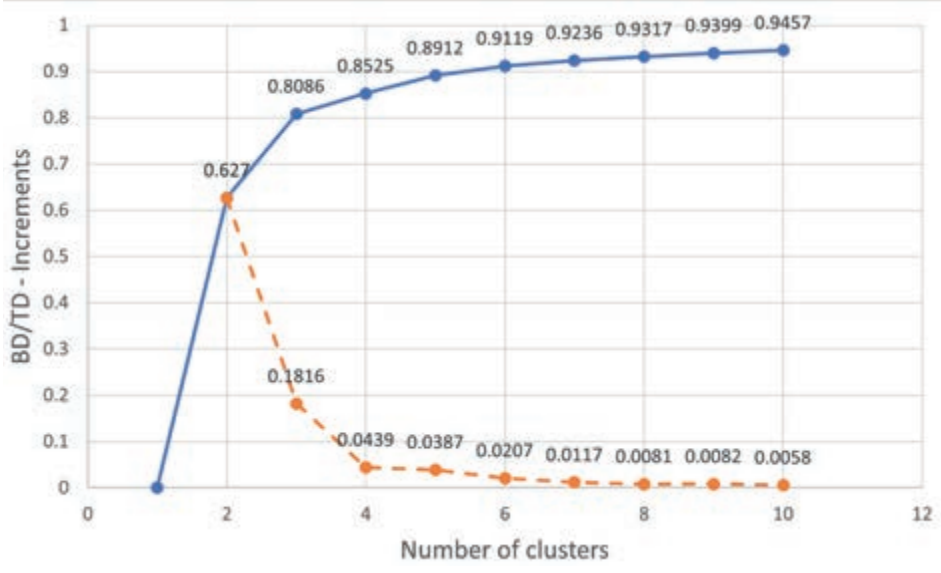


Figure 1: Quality of clusterisation (based on the coefficient of determination) as a function of the number of clusters. The blue line indicates the evolution of the amount of variance explained, while the dashed orange line portrays the increase in variance explained by each additional cluster.

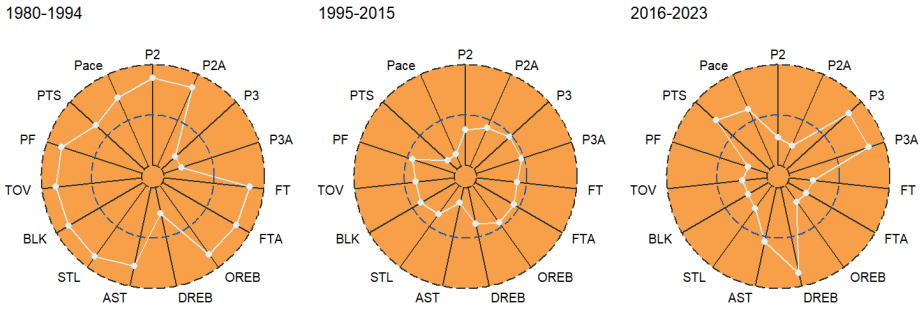


Figure 2: Standardised radial plots of the average profile of each era. The blue dashed line represents the overall (zero) mean of each stat after standardisation over all seasons.

according to the results, it is important to note that this could change based on how far back the seasons are considered and due to the ongoing third era. Interestingly, our clustering has yielded a subdivision similar to the one in a book by Rocha da Silva and Rodrigues[2] (a good read for those interested in deepening this topic), who combined principal component analysis and LASSO regression to their cluster analysis (compared to them, we added PF and Pace as attributes as well as the most recent NBA seasons). They proposed to refer to the three eras as the "Classic Era" (first era), the "Transitional Era" (second era) and the "Modern Era" (third era). In order to analyse the different characteristics of each cluster, one can take a look at the corresponding standardised radial plots (Figure 2).

In the Classic Era, the two-point shot seemed to be a large part of every team's playstyle, despite the three-point line having already been introduced. During this era, teams had above average stats in almost every category apart from three-point shooting and defensive rebounds. This could largely be due to the relatively high pace of the game, leading to more possessions and more opportunities to record certain stats. The Transitional Era seemed to present close to average stats in most categories, compared to the other two eras. A clear change was that teams started to rely more on three-point shots, as it became a reliable resource to win games. Meanwhile, the average pace tended to drop, explaining why most of the basic stats averages were lower than in the Classic Era. Lastly, the Modern Era was captivated by the high number of three-point attempts and makes, as well as points scored, defensive rebounds and overall pace. Overall, one can notice clear differences between the eras: the Classic Era mainly focused on close-distance shots, the Transitional Era started the shooting transition from the rim to the three-point line, and the Modern Era is characterised by an improved shot selection, shifting from less efficient two-point shots to more rewarding three-point shots.

In parallel to these developments, one could also notice that the enormous difference between the US national team and the rest of the world, as underlined by the large-margin victories of the American Dream Team during the Olympic Games of 1992, has clearly dropped, leading to much closer scores in recent international tournaments and even to Germany winning the 2023 FIBA Basketball World Cup. This makes the forthcoming Olympic Games tournament spicy and unpredictable!

As a statistician, it is rewarding to see that a well-known method like cluster analysis is capable of segmenting a complex game like basketball into distinct eras in a very satisfying manner (without gaps). Our work provides basketball fans a data-based foundation for their discussion by clearly separating the different playing eras.

When comparing players from different eras, it is important to keep the playstyle of that time in mind. With his unparalleled thirst for victory, Michael Jordan would certainly have trained his three-point shots with his usual zeal and, perhaps, matched Steph Curry's shooting performance. However, this remains a fiction whose reality we will never be able to check, not even by (Monte Carlo) computer simulations.

References

1. Zuccolotto, P. and Manisera, M. (2020) *Basketball Data Science, With Applications in R*. Boca Raton, FL: Chapman and Hall/CRC.

2. Rocha da Silva, J. V. and Rodrigues, P.C. (2021) The three Eras of the NBA regular seasons: Historical trend and success factors. *Journal of Sports Analytics*, **7**(4), 263–275.

Box 2: Basic basketball stats

***Points* (PTS). The number of points a player scores in a game. One point is received when scoring a free throw (FT), i.e., a free shot that you get after being fouled, two points when scoring a two-point field goal (P2), i.e., from inside the three-point line, and three points when scoring a three-point field goal (P3), i.e., from behind the three-point line. The number of attempted shots is consequently indicated as FTA, P2A and P3A.**

***Rebounds*. The number of times a player grabs the ball following a missed shot. They can be subdivided into offensive (OREB) and defensive (DREB) rebounds.**

***Assists* (AST). The number of times a player passes the ball to a teammate, who then scores.**

***Steals* (STL). The number of times a player gains possession of the ball from an opposing player.**

***Blocks* (BLK). The number of times a player blocks a shot of an opposing player.**

***Turnovers* (TO). The number of times a player loses possession of the ball to an opposing player.**

***Personal fouls* (PF). The number of times a player fouls an opposing player.**

***Pace*. An indication of the speed at which the game was played. It is proportional to the ratio between the total number of possessions of both teams (sequences of plays when they have the ball till it changes side) and the total number of minutes played by all players.**